

Компьютерная система контроля знаний по пунктуации русского языка

И.Н. Полякова, Е.А. Филимонова

Аннотация -- В работе предложена концепция компьютерной системы контроля знаний по пунктуации русского языка. Основу системы составляет алгоритм выявления в заданном предложении подходящего правила постановки запятой. Упомянутый алгоритм использует полученную на основании специальной информации, извлекаемой из предложения, формальную классификацию правил постановки запятых.

Использование в электронной переписке большого количества сокращений и упрощенных слов привело фактически к потере навыков грамотного написания текстов. Зачастую люди допускают ошибки в применении правил русского языка, а ведь правила являются не только показателем степени развития человека как личности, порой они играют важную роль для точного понимания написанного, как в известном примере “Казнить нельзя помиловать”. Проблему, порожденную ИТ-технологиями, можно, хотя бы частично, решить силами этих самых технологий, формализовав процесс выявления подходящих правил пунктуации.

Для целей автоматического выявления подходящего правила свод правил удобно разделить на группы. Классификация правил положена в основу модельной версии системы контроля знаний по пунктуации русского языка. По выбору пользователя система может функционировать в двух режимах: в режиме тестирования и в режиме обучения.

Классификация правил постановки запятой, первоначально разработанная для использования в программной системе контроля знаний, обладает также самостоятельной ценностью как наглядный путеводитель по своду (значительной части) правил русской грамматики.

Модельная версия системы контроля знаний по пунктуации русского языка проверена на достаточно представительном множестве примеров, позволяющих судить о перспективности предложенного подхода.

Ключевые слова— автоматическая обработка текста, алгоритм, анализ предложения, классификация правил пунктуации.

I. ВВЕДЕНИЕ

С появлением современных устройств связи потребность в устной речи сократилась. Вместе с тем, использование в электронной переписке большого количества сокращений и упрощенных слов привело фактически к потере навыков грамотного написания текстов. Зачастую люди допускают ошибки в применении правил русского языка [1], [2], а ведь правила являются не только показателем степени развития человека как личности [3], порой они играют важную роль для точного понимания написанного, как в известном примере “Казнить нельзя помиловать”.

Проблему, порожденную ИТ-технологиями, можно, хотя бы частично, решить силами этих самых технологий, формализовав процесс выявления подходящих правил пунктуации. Настоящая работа посвящена вопросам классификации и формализации правил постановки запятой [4], [5], [6] в предложении на русском языке (п.2), а также вопросам разработки модельной версии системы обучения пользователя этим правилам (п.3).

II. КЛАССИФИКАЦИЯ ПРАВИЛ ПУНКТУАЦИИ

Отправной точкой формализации правил постановки запятой является разбор предложения на составляющие компоненты. Традиционно различают три вида разбора: морфологический, синтаксический и семантический. Морфологический анализ позволяет определить часть речи, число, род, падеж заданного слова и приводит его к начальной форме [7], [8]. Синтаксический анализ позволяет выявить основные синтаксические конструкции и разбить заданное предложение на лексемы [9]. Семантический анализ позволяет выявить взаимосвязи между компонентами заданного предложения. В современных программных системах обработки текстов, как правило, используются сторонние модули, реализующие перечисленные виды анализа [10], [11], [12].

В русском языке имеется много правил, касающихся постановки запятой в предложении [1], [2], [13]. Для целей автоматического выявления подходящего правила свод правил удобно разделить на две группы:

1. Правила без вспомогательных слов и выражений, в том числе, правила, основанные на полном анализе предложения или анализе отдельных его частей;

Статья получена 13 декабря 2019.

Полякова И.Н., Московский государственный университет имени М.В. Ломоносова (email: polyakova@cs.msu.ru)
Филимонова Е.А., магистрант, Московский государственный университет имени М.В. Ломоносова (email: filkate0903@rambler.ru)

2. Правила со вспомогательными словами и выражениями:

2.1 правила, выявляющиеся по конкретному слову;

2.2 правила, выявляющиеся по группе слов или словосочетаниям.

Первую группу составляют правила, при проверке которых не используются вспомогательные слова и выражения. К таким относятся правила, основанные на полном анализе предложения или анализе отдельных его частей. К первой группе, в частности, относятся правила постановки запятой

- между простыми предложениями в составе сложного: «На небе сверкнула молния, и раздался гром»; или

- между однородными членами предложения без союзов и союзных слов: «На склонах растут сосны, кедры, лиственницы, березы».

Вторую группу составляют правила, при проверке которых используются вспомогательные слова и выражения. Особое место во второй группе занимают правила, подгруппы 2.1, которые однозначно определяются по одному слову. Типичными представителями этой подгруппы являются правила постановки запятой перед союзом «а» или перед союзом «но». Правила подгруппы 2.2 содержат более сложные конструкции, состоящие из двух и более слов. Для этих правил необходимо анализировать соседние слова с целью выявления:

- неделимых связей: «в то время как», «перед тем как», «подобно тому как», «не что иное, как», «все что угодно», «будь что будет», «ввиду того что» и пр.;

- связей, основанных на ближайшем окружении: «потому что», «если что», «как один» и пр.;

- связей, основанных на наличии или отсутствии в предложении слов: «что, когда ...», «что когда ... , то» и пр.

Для выявления всех правил постановки запятой используются результаты морфологического и синтаксического анализов, а для правил подгруппы 2.2 существенно используются также результаты семантического анализа.

III. СИСТЕМА КОНТРОЛЯ ЗНАНИЙ ПО ПУНКТУАЦИИ РУССКОГО ЯЗЫКА

Классификация правил положена в основу модельной версии системы контроля знаний по пунктуации русского языка. По выбору пользователя система может функционировать в двух режимах: в режиме тестирования и в режиме обучения. В режиме тестирования пользователь задает свой вариант расстановки запятых для некоторого предъявленного системой предложения, для которого верный вариант расстановки знаков препинания заранее известен (Рис. 1). При этом используется следующий алгоритм распознавания правил постановки запятой в предложении на русском языке:

Шаг 1. Перевод слов во внутреннее представление.

Шаг 2. Поиск проблемной позиции в предложении.

Шаг 3. Анализ меток и выявление связей.

Шаг 4. Выявление конкретного правила.

Шаг 5. Демонстрация формулировки правила с примером.

На шаге 1 некоторые слова и словосочетания предварительно помечаются метками, позволяющими перейти к обобщенному представлению предложения, в котором, с одной стороны, сохраняется информация для применения правил расстановки запятых, а, с другой стороны, устранена несущественная детализация.

На шаге 2 выявляется позиция в предложении, где пользователь допустил ошибку: поставил или пропустил запятую. Для выявления проблемной позиции исходное предложение сравнивается с его верным вариантом.

На шаге 3 проводится анализ меток и выявление связей слов или словосочетаний. При этом учитывается, что в русском языке существуют неделимые связки с точки зрения правил расстановки запятой в предложении.

На шаге 4 находится конкретное правило, в применении которого была допущена ошибка. При этом используется классификация правил. Первым несопадением может быть лишняя запятая. В этом случае алгоритм выдает позицию и пытается определить правило, которое не было соблюдено при постановке запятой. Если такое правило находится, то шаг 4 завершается. В противном случае считается, что запятая поставлена в месте, совершенно непредсказуемом, скажем, между словами, составляющими неразрывное словосочетание. Поскольку такая ошибка может возникнуть случайно, то алгоритм продолжает сравнивать верное и исходное предложения. Если сравнение заканчивается безрезультатно, то алгоритм сообщает пользователю о наличии ошибки без указания конкретного правила, по которому допущено несопадение.

На шаге 5 соответствующее правило выдается пользователю. Пользователь получает формулировку правила с примером.

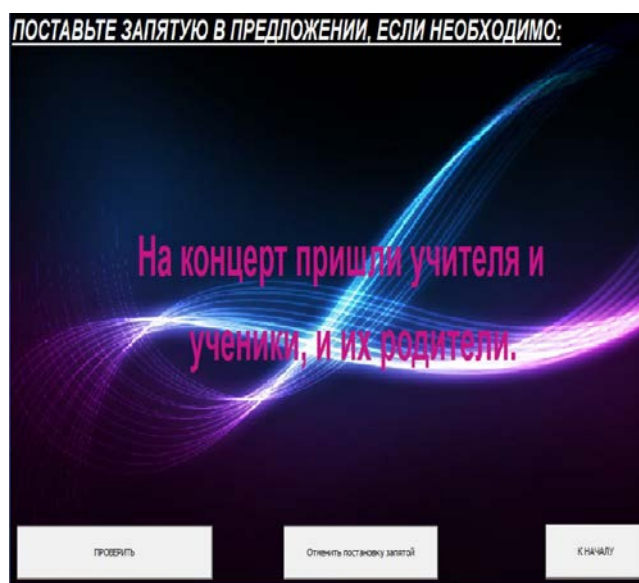


Рис. 1. Пример работы в режиме тестирования. Приведем еще некоторые скриншоты экранов при работе предложенной системы (Рис. 2 и Рис. 3).

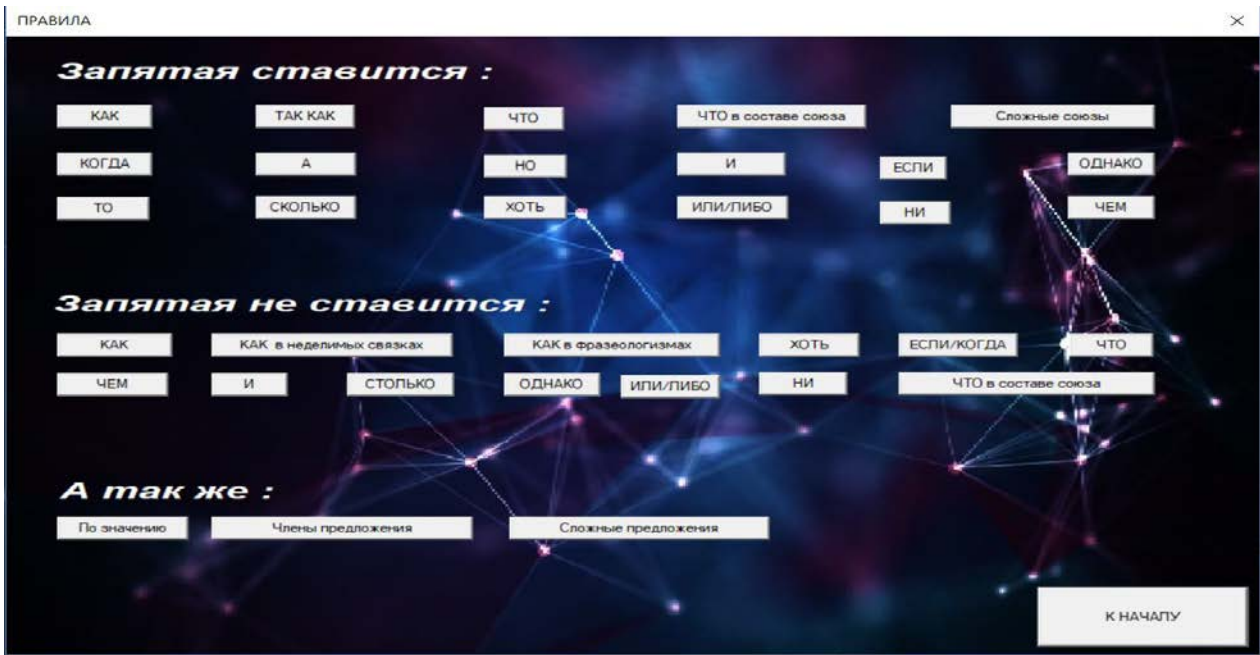


Рис. 2. Вид на сводку правил, доступный пользователю предложенной системы

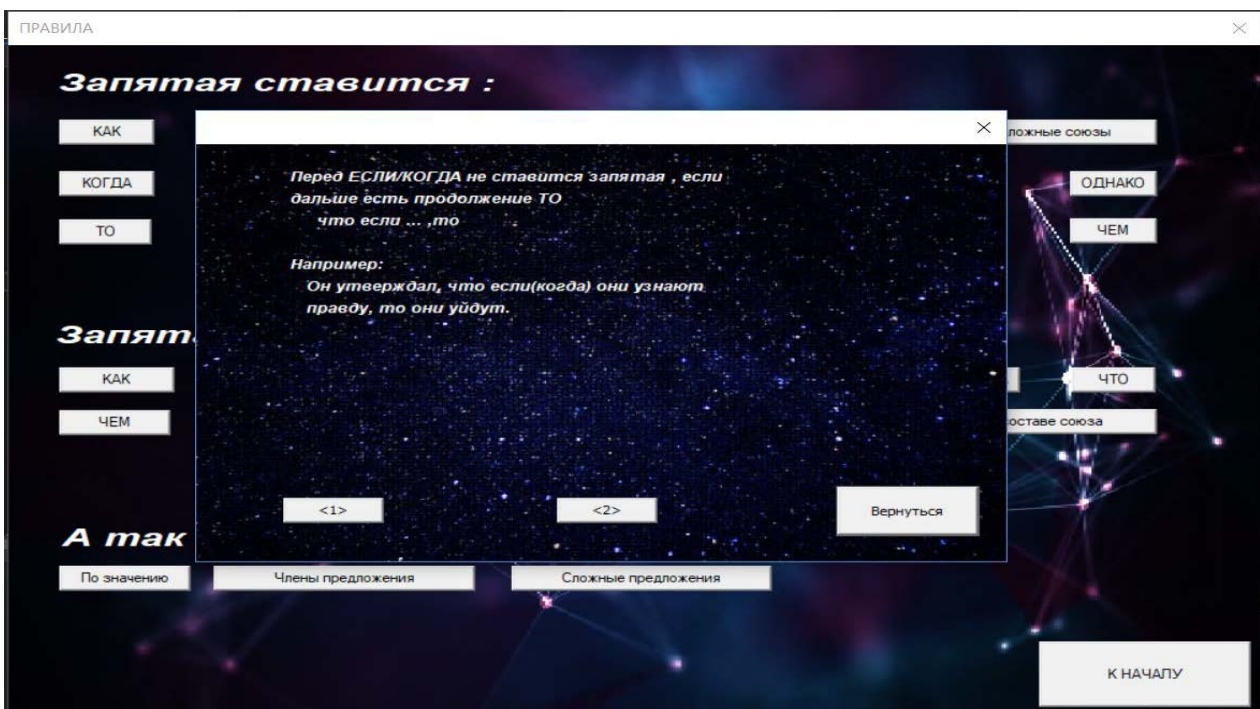


Рис. 3. Выборка конкретного правила с соответствующим примером

Модельная версия системы контроля знаний по пунктуации русского языка проверена на достаточно представительном множестве примеров, позволяющих судить о перспективности предложенного подхода.

IV ЗАКЛЮЧЕНИЕ

Классификация правил постановки запятой, первоначально разработанная для использования в программной системе контроля знаний, обладает также самостоятельной ценностью как наглядный путеводитель по своду (значительной части) правил русской грамматики.

Практически любой естественный язык в любом своем разделе не допускает простой алгоритмизации. Всегда найдется некоторое количество примеров, “не подчиняющихся” алгоритму и предполагающих отдельную обработку. Эксперименты с модельной реализацией описанной программной системы контроля знаний по пунктуации позволяет надеяться, что в деле грамотной постановки запятых количество исключений будет незначительным.

БЛАГОДАРНОСТИ

Авторы выражают благодарность профессору С.Ю.Соловьеву за ценные замечания и полезные советы, сделанные в ходе подготовки данной статьи.

БИБЛИОГРАФИЯ

- [1] Розенталь Д.Э. Справочник по русскому языку. Орфография и пунктуация. М.: Оникс, 2009. 368 с.
- [2] Валгина Н.С. Актуальные проблемы современной русской пунктуации. М.: Высш. шк., 2004. 259 с.
- [3] Маслова В.А. Введение в когнитивную лингвистику. М.: ФЛИНТА, 2018. 296 с.
- [4] Розенталь Д.Э., Голуб И.Б., Теленкова М.А. Современный русский язык. М.: Айрис-пресс, 2010. 448 с.
- [5] Великова Л.В. Русский язык для старшеклассников и абитуриентов. М.: МЦНМО, 2016. 344с.
- [6] Золотова Г.А., Онипенко Н.К., Сидорова М.Ю. Коммуникативная грамматика русского языка // М.: Институт русского языка РАН имени В.В.Виноградова, 2004. 544 с.
- [7] Арутюнова Н.Д. Проблемы морфологии и словообразования. М.: Языки славянских культур, 2007. 420 с.
- [8] Рябушкина С.В. Морфология современного русского языка: практикум. М.: Флинта: Наука, 2009. 256 с.
- [9] Смирнов И.В., Шелманов А.О. Семантико-синтаксический анализ естественных языков. Обзор методов синтаксического и семантического анализа текстов // Искусственный интеллект и принятие решений. 2013. № 1, с. 41-54.
- [10] Мальковский М.Г., Грацианова Т.Ю., Полякова И.Н. Прикладное программное обеспечение: системы автоматической обработки текстов. М.: МГУ, 2000. 52 с.
- [11] Автоматическая обработка текста. URL: <http://www.aot.ru> (дата обращения: 09.12.2019).
- [12] Боярский К.К. Введение в компьютерную лингвистику. СПб: НИУ ИТМО, 2013. 72 с.
- [13] Академия наук СССР, Министерство Высшего образования СССР, Министерство просвещения РСФСР. Правила русской орфографии и пунктуации. М.: 1956. 161 с.

Computer system of knowledge control on punctuation of Russian language

Irina Polyakova, Ekaterina Filimonova

Abstract— The paper proposes a concept of the computer system for controlling knowledge of punctuation in the Russian language. The system is based on the algorithm for the identifying a suitable comma rule in the given sentence. The mentioned algorithm uses a formal classification of comma rules derived from the special information extracted from the sentence.

The use of a large number of the abbreviations and the simplified words in the electronic correspondence has actually led to the loss of skills of literate writing. Often people make mistakes in the application of the rules of the Russian language, but the rules are not only an indicator of the degree of human development as individuals, sometimes they play an important role for the accurate understanding written, as in the famous example, "Execute not pardon". The problem generated by IT-technologies can be solved, at least partially, by means of these technologies, formalizing the process of the identifying suitable punctuation rules.

For the purpose of the automatically identifying suitable rule, it is convenient to divide the rulebook into some groups. The classification of the rules is the basis of the model version of the knowledge control system for punctuation of the Russian language. At the user's choice, the system can operate in two modes: test mode and training mode.

The classification of the comma placement rules was originally developed for the use in the knowledge control program system, but also it can be considered as the independent guide for a big part of the rules of the Russian grammar.

The model version of the knowledge control system for punctuation of the Russian language was tested on a fairly representative set of the examples that allow us to judge the prospects of the proposed approach.

Keywords — automatic text processing, algorithm, analysis of offers, classification of the rules of punctuation.

References

- [1] Rosenthal D. E. Guide of the Russian language. Spelling and punctuation. Moscow: Onyx, 2009. 368 p. (In Russ.)
- [2] Valgina N. S. Actual problems of modern Russian punctuation. M.: Higher school., 2004. 259 p. (In Russ.)
- [3] Maslova V. A. Introduction to cognitive linguistics. Moscow: Flinta, 2018. 296 p. (In Russ.)
- [4] Rosenthal D. E., Golub I. B., Telenkova M. A. Modern Russian language. Moscow: Iris press, 2010. 448 p. (In Russ.)
- [5] Velikova L. V. Russian language for high school students and applicants. Moscow: MCCME, 2016. 344 p. (In Russ.)
- [6] Zolotova G. A., Onipenko N. K., Sidorova M. Y. Communicative grammar of the Russian language. // Russian language Institute of the Russian Academy of Sciences named after V. V. Vinogradov, Moscow, 2004. 544 p. (In Russ.)
- [7] Arutyunova N. D. Problems of morphology and word formation. Moscow: Languages of Slavic cultures, 2007. 420 p. (In Russ.)
- [8] Ryabushkina S. V. Morphology of the modern Russian language: practical. Moscow: Flinta: Nauka, 2009. 256 p. (In Russ.)
- [9] Smirnov I. V., Shelmanov A. O. Semantic-syntactic analysis of natural languages. Review of methods of syntactic and semantic analysis of texts //

Artificial intelligence and decision-making . 2013. No. 1, p. 41-54. (In Russ.)

[10] Malkovsky M. G., Grazianova T. Yu., Polyakova I. N. Applied software: automatic text processing systems. Moscow: MSU, 2000. 52 p. (In Russ.)

[11] Automatic text processing . URL: <http://www.aot.ru> (date of issue: 09.12.2019) . (In Russ.)

[12] Boyarsky K. K. Introduction to computer linguistics. SPb: National research University of information technology, mechanics and optics, 2013. 72 p. (In Russ.)

[13] Academy of Sciences of the USSR, Ministry of Higher education of the USSR, Ministry of education of the RSFSR. Rules of Russian spelling and punctuation. Moscow: 1956. 161 p. (In Russ.)

Irina N. Polyakova, Associate Professor, Department of Algorithmic Languages, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1 Leninskie Gory, Moscow 119991, Russia), PhD in Physico-mathematical sciences, ORCID: <http://orcid.org/0000-0003-1432-4906>, polyakova@cs.msu.ru

Ekaterina A. Filimonova, Master's Degree student, Department of Algorithmic Languages, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1 Leninskie Gory, Moscow 119991, Russia), ORCID: <http://orcid.org/0000-0003-1155-4386>, filkate0903@rambler.ru